DOCUMENT RESUME

ED 478 066 TM 035 006

AUTHOR	Rupp, Andre A.; Zumbo, Bruno D.
TITLE	Bias Coefficients for Lack of Invariance in Unidimensional IRT Models.
PUB DATE	2003-00-00
NOTE	30p.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Item Response Theory; *Mathematical Models; *Statistical Bias
IDENTIFIERS	*Invariance; *Unidimensionality (Tests)

ABSTRACT

The feature that makes item response theory (IRT) models the models of choice for many psychometric data analysts is parameter invariance, the equality of item and examinee parameters from different populations. Using the well-known fact that item and examinee parameters are identical only up to a set of linear transformations specific to the functional form of a given IRT model, violations of these transformations for unidimensional IRT models are algebraically investigated and coefficients are derived for some violations. Since a lack of invariance constitutes item parameter drift (IPD) at the individual item level or item-set level, the magnitude and types of biases introduced by IPD along with their impact on examinee true scores can be algebraically derived, and these connections are demonstrated with results from a recently published simulation study (C. Wells, M. Subkoviak, and R. Serlin, 2002). This paper faciliates a deeper understanding of different types of lack of parameter invariance and their practical consequences for decision making through a framework that combines analytical, numerical, and visual perspectives on parameter invariance as a fundamental property of measurement. An appendix provides bias coefficients. (Contains 6 figures and 21 references.) (Author/SLD)

Running head: BIAS COEFFICIENTS FOR LACK OF INVARIANCE

Bias Coefficients for Lack of Invariance in Unidimensional IRT Models

André A. Rupp and Bruno D. Zumbo

University of British Columbia

Address Correspondence to:

André A. Rupp
The University of Ottawa
Faculty of Education
145 Jean-Jacques Lussier Street
Ottawa, Ontario, K1N 6N5
CANADA

e-mail: aarupp@interchange.ubc.ca

Abstract

The feature that makes item response theory (IRT) models the models of choice for many psychometric data analysts is *parameter invariance*, the equality of item and examinee parameters from different populations. Using the well-known fact that item and examinee parameters are identical only up to a set of linear transformations specific to the functional form of a given IRT model, violations of these transformations for unidimensional IRT models are algebraically investigated and bias coefficients are derived for some violations. Since a lack of invariance constitutes item parameter drift (IPD) at the individual item level or item-set level, the magnitude and types of biases introduced by IPD along with their impact on examinee true scores can be algebraically derived and these connections are demonstrated with results from a recently published simulation study (Wells, Subkoviak, & Serlin, 2002). This paper facilitates a deeper understanding of different types of *lack of parameter invariance* and their practical consequences for decision-making through a framework that combines analytical, numerical, and visual perspectives on parameter invariance as a fundamental property of measurement.

Bias Coefficients for Lack of Invariance in Unidimensional IRT Models

Item response theory (IRT) is one of the most popular current methodological frameworks for modeling response data from assessments. It is used directly in computer adaptive testing, cognitively diagnostic assessment, and test equating among other applications (e.g., Hambleton, Swaminathan, & Rogers, 1991; Junker, 1999; Kaskowitx & de Ayala, 2001). Furthermore, output from IRT models has more recently been incorporated into hierarchical regression models for multilevel data (e.g., Adams, Wilson, & Wu, 1997; Fox & Glas, 2001). The versatility of IRT models has made them the preferred tool of choice for many psychometric modelers, but beyond the flexibility of IRT models it is the often misunderstood feature of *parameter invariance* that is frequently cited in introductory or advanced texts as one of their most important characteristics (e.g., Hambleton & Jones, 1993; van der Linden & Hambleton, 1997; Hambleton et al., 1991; Lord, 1980). Since invariance relates to generalizability across contexts, parameter invariance in IRT models allows for the generalizability of inferences across context and thus constitutes a fundamental property of measurement.

In this paper, the mathematical formalization of parameter invariance is used to develop a framework for algebraic, numeric, and visual investigations of biases introduced by different types of *lack of invariance (LOI)*. The derivations in this paper are presented to clarify facets and implications of parameter invariance for a broad and more applied audience. In the term 'parameter invariance', the *parameters* referred to are both the set of *item parameters* and the set of *examinee parameters*. The word 'parameter' indicates that the term refers to *population quantities*, which are treated as fixed but unknown (in a frequentist framework) or random but unknown (in a Bayesian framework) and whose values are estimated with data collected within a random sampling framework. The word *invariance* indicates that parameter values are identical

in *separate* populations, which is commonly of concern when parameters are *estimated* repeatedly with different *calibration samples* that represent subsets of different populations of interest. Most importantly, parameter invariance denotes an *absolute ideal state* that holds only for *perfect model fit* and any discussion about whether there are "degrees of invariance" or whether there is "some invariance" are technically inappropriate (Hambleton et al., 1991). Moreover, the question of whether there is invariance in a given population is illogical as well as invariance requires *at least two populations or conditions* for parameter comparisons to be possible and meaningful.

The mathematical relationships that define parameter invariance are of course not novel per se and can be found, albeit often more cryptically, in other sources (e.g., Lord, 1980). In addition, work in score equating, differential item functioning (DIF), and item parameter drift (IPD) deals with LOI and the biases introduced thereby (e.g., Donoghue & Isham, 1998). However, the literature does not provide simple and widely accessible algebraic work on the conditions of parameter invariance and possible violations of these, which is why the work in this paper seeks to clarify many of the subtleties of parameter invariance for practitioners and theoreticians alike.

Mathematically, parameter invariance is a simple *identity* for parameters that are on the same scale. Yet the latent scale in IRT models is arbitrary so that unequated sets of model parameters are invariant only up to a set of linear transformations *specific* to a given IRT model. When *estimating* these parameters in *unidimensional* IRT models with *calibration samples*, this indeterminacy is typically resolved by requiring that the latent indicator θ be normally distributed with mean 0 and standard deviation 1 (i.e., $\theta \sim N(0,1)$). In *orthogonal multidimensional* IRT models, the latent scale indeterminacy implies that parameters are

identical up to an orthogonal rotation, a translation transformation, and a single dilution or contraction. When estimating these parameters with calibration samples, the indeterminacy is typically resolved by requiring that the multivariate latent indicator $\boldsymbol{\theta}$ be multivariate normally distributed with mean **0** and variance-covariance matrix **I** where **I** is the identity matrix of appropriate size (i.e., $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \mathbf{I})$), which is the multidimensional analogue to the unidimensional case (Davey, Oshima, & Lee, 1996; Li & Lissitz, 2000). Once estimated values of the parameters for different populations are available on their respective scales, it is of interest to determine the type of relationship that exists between them as a yardstick to assess whether the same IRT model is likely to hold in both populations (i.e., whether parameter invariance across the populations holds). However, the methods that are used to assess a LOI need to be carefully chosen as simple indices such as correlation coefficients may miss additive group level effects, for example (Rupp & Zumbo, 2002).

In this paper we use the term *bias coefficients* while acknowledging that the work 'bias' has a variety of different usages in the statistical and non-statistical literature. In textbooks of statistical inference, bias is generally defined as the difference between the expected value of an estimator and the quantity it is trying to estimate (Casella & Berger, 1990, p. 303). In the literature on differential item functioning (DIF), bias is sometimes referred to as an undesired differential functioning of items that is not attributable to ability differences on the latent dimensions the test is intended to measure. Under this operationalization bias produces an unfair advantage for one group of examinees over another as the examinees in both groups possess differing amounts of proficiency on the nuisance dimensions (Shealy & Stout, 1993). In this paper, the term bias coefficients is used to denote quantities that are derived from differences in model parameters due to IPD, because, if IPD goes undetected, the examinees are assigned a score that is different

from the one they should be correctly assigned if the drift were detected. As an additional point of clarification, all of the following equations involve *population* quantities only, because the focus of this paper is *not* the estimation of biases but the analytical derivation of the idealized population analogues. Circumventing the estimation process allows for discussions of what can be considered "best-case" and "worse-case" scenarios with any real data applications being instantiations of these cases.

To derive bias coefficients, consider the unidimensional two-parameter logistic (2PL) model for illustrative purposes where examinees are indexed by $i = 1,\ldots, I$, items are indexed by $j = 1,\ldots, J$, and $P_j(\theta_i)$ is the probability of examinee i responding to item j correctly as a function of the latent trait θ. The 2PL model can be written as follows (Hambleton, 1989):

$$P_j(\theta_i) = \frac{\exp(\alpha_j(\theta_i - \beta_j))}{1+\exp(\alpha_j(\theta_i - \beta_j))} \; ; \alpha_j > 0 \, , -\infty < \beta_j, \theta_i < \infty$$

where α_j is the item slope or "item discrimination" parameter, β_j is the item location or "item difficulty" parameter, and θ_i is the latent predictor or "proficiency" variable. In the following parameters from a second population of interest are denoted by a prime ('); conceptually, *neither* population is considered more 'important' in any sense. Thus, they will not be semantically distinguished with terms such as 'reference' or 'focal' population as is done in, for example, the literature on differential item functioning (DIF; see Clauser & Mazor, 1998; Donoghue & Isham, 1998; Zumbo, 1999).

For parameters in the 2PL to be invariant in the populations of interest, one simply requires $\alpha_j' = \alpha_j$, $\beta_j' = \beta_j$, and $\theta_i' = \theta_i$ to hold *jointly for all items and examinees* that are relevant to the practical context at hand if the parameters are *linked* onto the same scale. Due to the

indeterminacy of the latent scale for θ, the above identities are equivalent to the following equations for *unlinked* scales:

$$\alpha_j^{'} = \delta^{-1}\alpha_j$$
$$\beta_j^{'} = \varepsilon + \delta\beta_j$$
$$\theta_i^{'} = \varepsilon + \delta\theta_i$$

where ε and δ are non-zero real numbers. Mathematically, parameters *fail* to be invariant if *at least one* of these equations does not hold for *at least one* item or examinee in the populations of interest depending on which parameters are investigated for invariance. The above equations represent *restrictive* kinds of linear transformations, which is why it is inappropriate to compare parameter estimates from different calibrations with indices that measure linear association only. Hence, considerations about invariance need to include considerations of item sets as well as of individual items (e.g., Donoghue & Isham, 1998, Zumbo, 2003). To understand the types of biases that are possible under a LOI, it is insightful to consider the impact of different violations of the conditions above on the response probabilities.

In generic terms, the linked parameters from the first and second population are related by $\alpha' = f(\alpha)$, $\beta' = g(\beta)$, and $\theta' = h(\theta)$ and are invariant only if the transformation functions $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ are *identity* functions *for all items and examinees*; otherwise, they fail to be invariant. For the sake of simplicity, the following examples of parameter invariance will be restricted to item parameters and will consider only *linear* transformation functions for α and β; the derivations for θ are very similar to those for β since the two parameters are on the same scale even though the meanings behind these invariance investigations are very different. Since each linear relationship is represented by a line with an intercept and a slope, there are three cases to consider for each item parameter.

For α_j we have

$$\alpha_j^{'} = \alpha_j + \tau_j \qquad \text{(I)}$$
$$\alpha_j^{'} = \omega_j \cdot \alpha_j \qquad \text{(II)}$$
$$\alpha_j^{'} = \omega_j \cdot \alpha_j + \tau_j \qquad \text{(III)}$$

where ω_j and τ_j are non-zero real numbers.

For β_j we similarly have

$$\beta_j^{'} = \beta_j + \kappa_j \qquad \text{(IV)}$$
$$\beta_j^{'} = \lambda_j \cdot \beta_j \qquad \text{(V)}$$
$$\beta_j^{'} = \lambda_j \cdot \beta_j + \kappa_j \qquad \text{(VI)}$$

where κ_j and λ_j are non-zero real numbers. Note that these six cases are *not* distinguishable for a given item. That is, if an item parameter value has drifted and only the drifted value is observed – as is generally the case when we work with estimates – then there is exactly one real-valued constant τ_j, one real-valued factor ω_j, and an infinite number of real-valued pairs $\{\tau_j, \omega_j\}$ that could have given rise to the transformed value. However, if a transformation applies to *sets* of items, a distinction between the above cases is crucial as the biases under different transformations are of different form and magnitude across the drifted items. The six basic cases (I) – (VI) lead to a total number of 15 cases if joint violations of invariance in α_j and β_j are considered. However, they will not all be described in detail because some cases are combinations of the six basic cases and follow logically from those. Hence, in the following section, only the six basic cases are used to express biases first on the *logit scale.* The section after that then shows how these biases can be translated into biases on the *probability scale* to clearly highlight their practical utility as differences in response probabilities and related true scores are the focus of practical decision-making.

Bias on the Logit Scale

For some cases, the biases that are introduced by violations (I) – (VI) can be compactly written with coefficients on the scale that is defined by the link function. The logit scale was chosen for analytical convenience but any other transformation with appropriate properties (e.g., the probit transformation) will technically work as well. We will present only the simple cases (I) and (IV) below and have collected the remaining four cases in the Appendix. For each case (a) the new relationship between the parameter values and (b) the introduced bias on the logit scale will be presented. The bias coefficients will then be interpreted but it is thus crucial to note that the interpretation is with respect to the *logit scale* and does not necessarily mirror the interpretation that would be appropriate on the probability scale. Since most practitioners are probably more interested in the implications of biases for response probabilities and test scores, these will be discussed in a later section and several biases on the *probability scale* will be interpreted there. The following description therefore primarily highlights succinctly the interrelationships between the parameter transformation function (i.e., the type of LOI) and the logit scale formulation of the two-parameter kernel.

Case (I) – Non-zero intercept for α'

For non-zero real numbers δ and τ_j,

(a) $\alpha_j' = \delta^{-1}(\alpha_j + \tau_j) = \delta^{-1}\alpha_j + \delta^{-1}\tau_j$

(b) $\text{logit}[P_j'(\theta_i')] = (\alpha_j + \tau_j)(\theta_i - \beta_j) = \alpha_j(\theta_i - \beta_j) + \tau_j(\theta_i - \beta_j) = \text{logit}[P_j(\theta_i)] + B_j^{\theta,\beta}$

where $B_j^{\theta,\beta} = \tau_j(\theta_i - \beta_j)$ is an *additive* bias coefficient whose absolute magnitude depends on the location difference θ_i - β_j and δ is the *global* transformation parameter – hence, no subscript – required to link scales. Therefore, for a given item, the introduced logit-scale bias is larger in absolute magnitude for an examinee whose ability is very different from the difficulty of the item

than for an examinee whose ability level is closer to the difficulty of the item. No bias exists for examinees whose ability level is identical to the item difficulty.

Case (IV) – Different intercept for β'

For non-zero real numbers ε, δ, and κ_j,

(a) $\beta_j' = \varepsilon + \delta(\beta_j + \kappa_j) = (\varepsilon + \delta\kappa_j) + \delta\beta_j$

(b) $\text{logit}[P_j'(\theta_i')] = \alpha_j(\theta_i - (\beta_j + \kappa_j)) = \alpha_j(\theta_i - \beta_j) - \alpha_j\kappa_j = \text{logit}[P_j(\theta_i)] + B_j^{\alpha}$

where $B_j^{\alpha} = -\alpha_j\kappa_j$ is an *additive* bias coefficient whose magnitude depends, for each item, on its discrimination parameter and ε and δ are the *global* transformation parameters – again, no subscript – required to link scales. Therefore, items with higher discrimination values will have a larger *logit-scale bias* independent of the location difference between examinee and item, which is actually an accurate description of the bias on the probability scale for this case as well.

In all cases it is clear that the biases result in differences in *item characteristic curves (ICCs)*, which equal differences in *response probabilities* for all or almost all examinees. But since the logit transformation is non-linear, the effects of biases on the logit and probability scales are different and the additivity of bias on the logit scale is not preserved on the probability scale. It is thus necessary to translate the *logit-scale biases* into *probability-scale biases*. The following section discusses the practical utility of the bias coefficients for the estimation of response probabilities and true scores and shows how these results are useful for the study of IPD.

Bias on the Probability Scale

It is possible to use the above formulations to analytically compute differences in response probabilities at the population level as is done empirically in studies of IPD for calibration samples (e.g., Wells et al., 2002; see also Donoghue & Isham, 1998). Conceptually, IPD is typically defined as the differential shift of item parameters over time (Goldstein, 1983), which is

often attributed to educational, technological, or cultural changes (Bock, Muraki, & Pfeiffenberger, 1988). Mathematically, it is readily seen that IPD represents LOI *at the item* level where IPD in either α or β leads to a change in the respective parameter value with the form of the exact transformation from α to α' or β to β' unknown. Hence, one way to represent IPD at the item level is

$$\alpha_j^{'} = \alpha_j + \tau_j \quad \text{(A)}$$
$$\beta_j^{'} = \beta_j + \kappa_j \quad \text{(B)}$$

where $-\alpha_j < \tau_j < \infty$, $-\infty < \kappa_j < \infty$ with the first inequality ensuring that $\alpha_j^{'} > 0$. In other words, all cases (I) – (VI) are cases of IPD but the simplest way to *simulate* drift and to analytically investigate it is by casting it as an *additive* formulation. Since graphical comparisons of ICCs are made on the probability scale it is helpful to translate the above statements into bias statements on that scale. To combine the discussion for both cases into one, consider a *general additive bias on the logit scale* where ϕ is any non-zero real number:

$$\text{logit}[P_j^{'}(\theta_i^{'})] = \alpha_j(\theta_i - \beta_j) + \phi = \text{logit}[P_j(\theta_i)] + \phi .$$

On the probability scale, this is written as

$$P_j^{'}(\theta_i^{'}) = \frac{\exp[\alpha_j(\theta_i - \beta_j)]\exp[\phi]}{1 + \exp[\alpha_j(\theta_i - \beta_j)]\exp[\phi]} = \frac{\exp[\alpha_j(\theta_i - \beta_j)]}{\exp[-\phi] + \exp[\alpha_j(\theta_i - \beta_j)]},$$

which can be compared to the response function with item parameters that have not drifted,

$$P_j(\theta_i) = \frac{\exp[\alpha_j(\theta_i - \beta_j)]}{1 + \exp[\alpha_j(\theta_i - \beta_j)]} .$$

A few basic algebraic steps result in the following relationships:

$$\phi < 0 \Rightarrow P_j^{'}(\theta_i^{'}) < P_j(\theta_i) \qquad \text{(R1)}$$
$$\phi = 0 \Rightarrow P_j^{'}(\theta_i^{'}) = P_j(\theta_i) \qquad \text{(R2)}$$
$$\phi > 0 \Rightarrow P_j^{'}(\theta_i^{'}) > P_j(\theta_i) \qquad \text{(R3)}$$

where the arrow denotes an implication. In other words, if the additive logit-scale bias is positive, the probability under the drifted parameters will be positively biased; if it is negative, it will be negatively biased; otherwise, the two probabilities will be identical. The relationships (R1) – (R3) are *not* equivalences, however, because differences in response probabilities can have many causes only one of which is an additive bias on the logit scale.

Consider the Wells et al. (2002) study for illustrative purposes. The authors simulated drift in the population values of the item difficulty and discrimination parameters in a 2PL. Only positive amounts of drift were considered and the effect of item parameter drift on the estimation of examinee ability parameters was estimated under 48 conditions: Test length (2 levels) × sample size (2 levels) × type of drift (3 levels) × number of drift items (4 levels). More specifically, if an item was selected to display item parameter drift, the authors increased either the discrimination parameter by .5 or the difficulty parameter by .4 or both simultaneously by .5 and .4 respectively.

It is immediately clear that increasing a difficulty parameter by some positive number leads to an ICC that is shifted to the right and that increasing a discrimination parameter by some positive number leads to an unchanged inflection point but a steeper slope. Yet, in addition to the conceptual understanding, it is possible to quantify these changes more precisely and the bias coefficients allow us to do just that. When the authors changed an item discrimination parameter value by .5, they introduced a bias of

$$B_j^{\theta,\beta} = \tau_j(\theta_i - \beta_j) = .5(\theta_i - \beta_j)$$

according to case (I). For drifted items, this results in ICC segments that are shifted upward for positive bias (R3), which occurs when $\theta_i > \beta_j$, ICC segments that are shifted downward for negative bias (R1), which occurs when $\theta_i < \beta_j$, and an identical ICC value for no bias (R2) at $\theta_i = \beta_j$. This pattern was observed (see Wells et al, 2002, Figure 1a, p. 80) and plotted with respect to the estimated true score, which is computed as the sum of the ICCs over all items in the test:

$$T(\theta_i) = \sum_{j=1}^{J} P_j(\theta_i)$$

The resulting curve that traces the true score as a function of the latent indicator θ is called the test characteristic curve (TCC) and it was seen that the overall shift in the TCC was relatively minimal, because only a few items exhibited drift in each of the design conditions in the study. This also stems from the fact that the differences in response probabilities are actually relatively minor. To illustrate this, let us formally denote the difference in response probabilities by Δ_{ij},

$$\Delta_{ij} = P_j(\theta_i) - P_j^{'}(\theta_i^{'}) \quad \text{with} \quad -1 \leq \Delta_{ij} \leq 1 \ .$$

Table A1 shows the Δ_{ij} values (cell entries) for an α-drift of .5 as a function of the original discrimination value of an item (row value) and the location difference between an examinee and an item on the θ scale, $\theta_i - \beta_j$ (column value). For example, take an item with an original discrimination value of .75 and an examinee whose location on the latent scale is .5 units above the location of the item (i.e., $\theta_i - \beta_j = .5$). The bias that gets introduced for this examinee on this item under drift of the discrimination parameter manifests itself in a difference in response probabilities of only $\Delta_{ij} = -.0586883$. In other words, the response probability under the drifted discrimination parameter is about 6% *higher* than under the non-drifted parameter. It can be seen

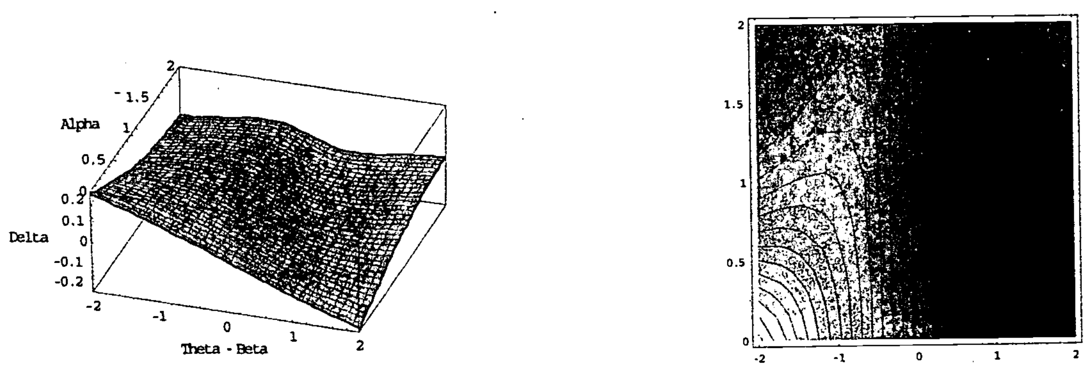
in this table that most Δ_{ij} values are between .05 and .10. In other words, between 10 to 20 items with an α-drift of this magnitude are necessary to result in a true score difference of only 1 point, a difference that would probably be considered rather trivial for most practical circumstances.

When the authors changed an item difficulty parameter value by .4, they introduced a negative bias of $B_j^{\alpha} = -\omega_j \alpha_j = -.4\alpha_j < 0$, according to case (III), where the inequality stems from the fact that item discrimination values are always positive. For drifted items, this results in ICCs that are shifted to the right according to (R1) independent of the values of θ_i and β_j, which was observed with again relatively minimal effects in terms of the TCC (see Wells et al., 2002, Figure 2a, p. 82). Again, the Δ_{ij} values for a variety of item discrimination parameters and location differences can be computed (see Table A2) and, again, most of the Δ_{ij} values for moderately discriminating items and moderate location differences are between .05 and .10 albeit some cases with higher values can be observed. Just as before this means that for the majority of cases between 10 to 20 items with a β-drift of this magnitude are required to produce a true-score change of 1 point, a relatively minor effect.

Finally, when the authors changed both the item discrimination parameter value by .5 and the item difficulty parameter value by .4, they introduced multiple biases. Even though conditions for when upward and downward shifts of the ICCs occur can be formally stated, those conditions are relatively cumbersome to present and are omitted here. In the study, the authors report that the TCCs cross at a value θ_0 where $\theta_0 > \beta_j$. The effects of the biases as seen in the TCCs were again relatively minimal for most values of θ but now started to increase in magnitude for specific sub-regions of the θ space compared to the previous scenarios (see Wells et al., 2002, Figure 3a, p.83). Table A3 shows the Δ_{ij} values for this scenario and it can be seen that these are

still often between .05 and .10 but that now there are also quite a few values in the range of .10 to .15 with some even reaching .20. Thus, even though between 10 and 20 items are required to produce a true score difference of 1 point for many cases under this joint α- and β-drift, only around 5 to 7 items are required in other cases.

All three scenarios show that when the pattern of introduced biases is expressed with respect to response probabilities it appears quite complex due to the curvature and asymptotic behavior of the ICCs. It is possible to plot the Δ_{ij} values to illustrate this complex behavior more closely. Figure 1 shows the Δ_{ij} surface and contour plots for the α-drift of .5, item discrimination values of non-drifted items between 0 and 2, and location differences between –2 and 2 to match the structure of Table A1 while utilizing more grid points. Note that for the surface plot Δ_{ij} is labeled 'Delta', the location difference is labeled 'Theta-Beta', and the non-drifted discrimination values are labeled 'Alpha'. Furthermore, the orientation of the contour plot matches the orientation of the surface plot so that the horizontal axis represents the location difference values, the vertical axis represents the item discrimination values, and the contour lines and shades represent the Δ_{ij} values with lighter shades corresponding to higher Δ_{ij} values and darker shades corresponding to lower Δ_{ij} values.



(a) Surface Plot for α-drift

(b) Contour Plot for α-drift

Figure 1. Surface and contour plots of Δ_{ij} for $\alpha' = \alpha + .5$.

To understand these plots, note that when an item discrimination parameter drifts the slope of the ICC for the item with the drifted parameter is steeper, which results in *increasing* differences in response probabilities in *both* directions from the inflection point for some range of θ values followed by *decreasing* differences as the original ICC and the ICC under drift approach their asymptotes. As an example of this behavior, Figure 2 shows a plot of a drifted item with original discrimination value $\alpha = 1$, discrimination value $\alpha' = \alpha + .5 = 1.5$ after drift, and difficulty value $\beta = 0$. Note that the latent trait θ is labeled 'Theta' and that $P_j(\theta_i)$ is labeled 'Probability'.

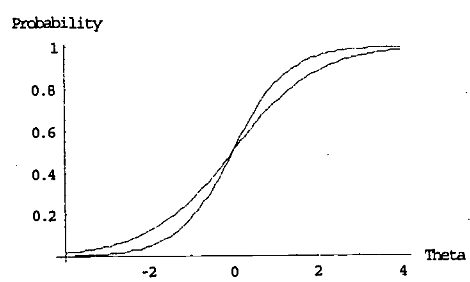
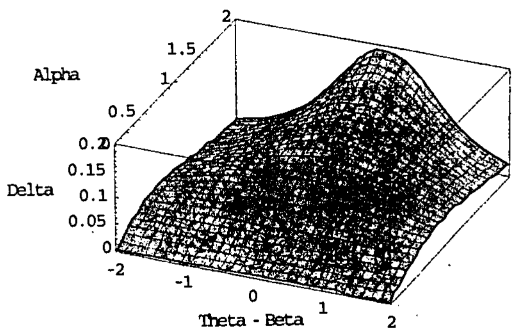


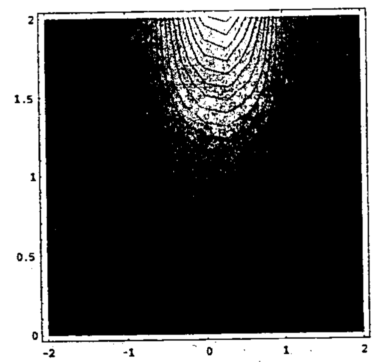
Figure 2. ICCs for item with α-drift of .5.

The differences are *positive* to the left of the inflection point and *negative* to the right of the inflection point as a result of how Δ_{ij} was defined here. The surface and contour plots of Figure 3 show graphically how differences are *largest* in absolute magnitude for the *least* discriminating items and *smallest* for the *most* discriminating items for the range of location differences considered. This makes sense, because if the slope of an ICC is already rather steep without drift present (i.e., when an item is already highly discriminating) then a further increase in slope will have relatively little impact on response probability differences. This implies that if items are of at least reasonable discriminatory power for a given population (e.g., if they have an α_j value of at least 1) biases are not as extreme.

Figure 3 shows the Δ_{ij} surface and contour plots for the β-drift of .4, item discrimination values of non-drifted items between 0 and 2, and location differences between –2 and 2 to match the structure of Table A2 while again utilizing more grid points. The labeling corresponds to that of Figure 1.



(a) Surface Plot for β-drift



(b) Contour Plot for β-drift

Figure 3. Surface and contour plots of Δ_{ij} for $\beta' = \beta + .4$.

This shows that when an item difficulty parameter drifts, the effect is *asymmetric* with respect to the inflection point of the ICC. Figure 4 shows a plot of a drifted item with original difficulty value $\beta = 0$, drifted difficulty value $\beta' = \beta + .4 = .4$, and discrimination value $\alpha = 1$ to illustrate this behavior. The labeling corresponds to that of Figure 2.

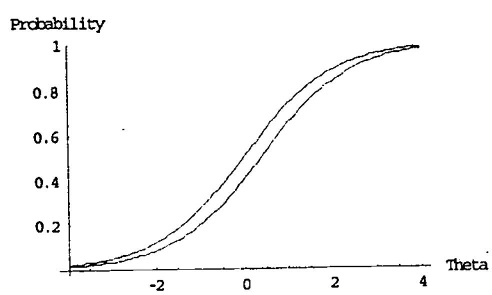
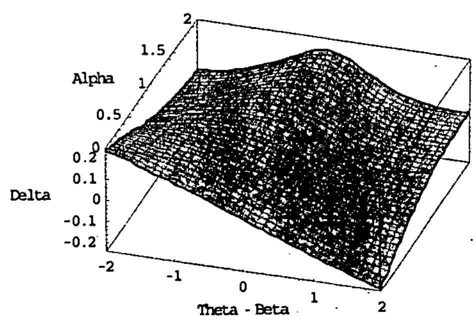


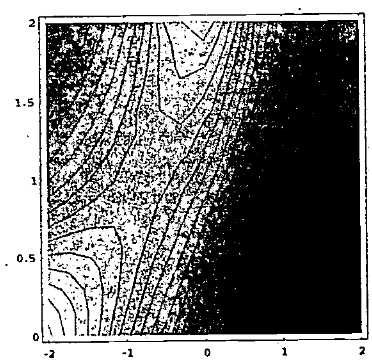
Figure 4. ICCs for item with β-drift of .4.

As previously seen in the surface and contour plots of Figure 3, the difference in response probabilities gets larger in absolute magnitude as the discrimination value gets larger and items with higher discrimination values have a higher bias for smaller location differences.

Finally, Figure 5 shows the Δ_{ij} surface and contour plots for a joint α-drift of .5 and β-drift of .4 for item discrimination values between 0 and 2, and location differences between –2 and 2 to match the structure of Table A3 while again utilizing more grid points. The labeling corresponds to that of Figure 1 and Figure 3.



(a) Surface Plot for joint α- and β-drift



(b) Contour Plot for joint α- and β-drift

Figure 5. Surface and contour plots of Δ_{ij} for $\alpha' = \alpha + .5$ and $\beta' = \beta + .4$.

This plot shows the complex effects that both drift types have on the difference in response probabilities and one can readily identify characteristics of the previous two cases. For example, note the almost linear difference values for poorly discriminating items in the location difference range considered here due to flat ICCs and the pronounced spike in difference values for highly discriminating items similar to the cases before. As an example of the complex behavior of the Δ_{ij} values, Figure 6 shows the ICC of an item with original parameter values $\beta = 0$, $\alpha = 1$, and the ICC of the same item with drifted parameter values $\beta' = .4$ and $\alpha' = 1.5$. The values were chosen to match the effects shown in Figure 2 and Figure 4 and the labeling is identical to these figures as well:

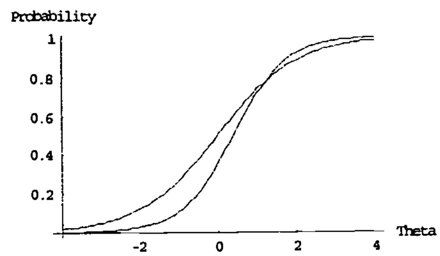


Figure 6. Plot of ICCs for item with joint α-drift of .5 and β-drift of .4.

These complex relationships raise the issue of what kind of discrimination values and location differences are typically observed in practice. It seems clear that extreme differences of, say, ± 2 or 3 units, can be observed in many practically relevant cases if test data are collected with item and examinee population subsets that yield a wide range of item and examinee parameter values. Indeed, a good test often consists of items with a wide variety of difficulty levels and a moderate range of discrimination values and is typically given to examinees with a wide range of ability levels with the implicit hope that item and examinee properties are well captured by the chosen model. Whether or not the intersection of a given examinee with a given item results in a large bias under drift of some parameter cannot be generally answered, however, and depends on the type and magnitude of drift.

Conclusions

This paper also underscored that parameter invariance is an ideal state that is technically violated if at least one identity condition does not hold for at least one examinee or item. Violations can be of any kind but three linear non-identity transformations were considered and the biases introduced on the logit scale by this LOI were represented with bias coefficients whenever possible. The bias coefficient framework primarily serves to highlight the dependencies of different types of bias on model parameters that have not drifted and it allows one to quickly gauge the severity of biases on the logit and probability scales. From a practical

viewpoint, the different perspectives taken here allow one to compute and visualize different biases directly, without having to resort to simulation studies or real data sets, which can easily be done for a variety of different conditions. The framework can thus be used to cleanly assess the impact certain biases have on the response probabilities and examinee true scores; any real-life data set will be a mixed bag of different biases that falls somewhere between the clean analytical extremes. Most importantly, this paper and other research suggest that IRT models inferences about examinees are relatively robust toward moderate amounts of IPD across a wide range of theoretical conditions. It is hoped that this paper contributes to the on-going process of clarifying what is meant by parameter invariance and to demystify its status, which is often misperceived as a "mysterious" property that all IRT models seem to possess by definition across an almost infinite range of populations and conditions. If sound theoretical discussions about scientific generalizability are desired, this paper shows that the mathematical foundations of parameter invariance as a fundamental property of measurement cannot be ignored.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. Journal of Educational and Behavioral Statistics, 22, 47-76.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. Journal of Educational Measurement, 25, 275-285.

Casella, G., & Berger, R. L. (1990). Statistical inference. Belmont, CA: Duxbury Press.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning items. Educational Measurement: Issues and Practice, 17, 31-45.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. Applied Psychological Measurements, 20, 405-416.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. Applied Psychological Measurement, 22, 33-51.

Fox, J., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika, 66, 271-288.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. Journal of Educational Measurement, 20, 369-377.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational Measurement (pp. 147 – 200). New York: Macmillan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice, 12, 38-47.

Junker, B. W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment. Unpublished manuscript. Available online at http://www.stat.cmu.edu/~brian/nrc/cfa

Kaskowitz, G. S., & de Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. Applied Psychological Measurement, 25, 39-52.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. Applied Psychological Measurement, 24, 115-138.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Rupp, A. A., & Zumbo, B. D. (2002). How to quantify and report whether parameter invariance holds: When Pearson correlations are not enough. Manuscript submitted for publication.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58, 159-194.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. Applied Psychological Measurement, 26, 77-87.

van der Linden, W. J., & Hambleton, R. K. (1997). Handbook of modern item response theory. New York: Springer-Verlag.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. Language Testing, 20, 136-147.

Appendix

A1 – Values of Δ_{ij} for $\alpha' = \alpha + .5$

Discrimination Parameter α_j	Location Difference θ_i - β_j: -2	-1.5	-1	-.50	0	.5	1	1.5	2
.25	0.195115	0.162248	0.117002	0.0614572	0	- 0.0614572	- 0.117002	- 0.162248	- 0.195115
.50	0.149738	0.138396	0.108599	0.0602828	0	- 0.0602828	- 0.108599	- 0.138396	- 0.149738
.75	0.106567	0.112121	0.0981212	0.0586883	0	- 0.0586883	- 0.0981212	- 0.112121	- 0.106567
1	0.071777	0.0870761	0.0865159	0.0567194	0	- 0.0567194	- 0.0865159	- 0.0870761	- 0.071777
1.25	0.0465459	0.0654175	0.0746529	0.0544302	0	- 0.0544302	- 0.0746529	- 0.0654175	- 0.0465459
1.5	0.0294397	0.0479236	0.0632226	0.0518799	0	- 0.0518799	- 0.0632226	- 0.0479236	- 0.0294397
1.75	0.0183253	0.0344607	0.0526977	0.04913	0	- 0.04913	- 0.0526977	- 0.0344607	- 0.0183253
2	0.0112934	0.0244485	0.0433447	0.0462413	0	- 0.0462413	- 0.0433447	- 0.0244485	- 0.0112934

Note: $\Delta_{ij} = P_j(\theta_i) - P_j'(\theta_i')$.

A2 – Values of Δ_{ij} for $\beta' = \beta + .4$

Discrimination Parameter α_j	Location Difference θ_i - β_j								
	-2	-1.5	-1	-.50	0	.5	1	1.5	2
.25	0.023197	0.0238999	0.0244411	0.0248045	0.0249792	0.0249597	0.0247467	0.0243466	0.0237717
.50	0.0374662	0.0419365	0.0457284	0.0484627	0.049834	0.0496791	0.0480168	0.0450431	0.0410841
.75	0.0405745	0.0512061	0.0615962	0.0699552	0.0744425	0.0739254	0.0685395	0.0596183	0.0490497
1	0.0360302	0.052317	0.0711253	0.0884902	0.0986877	0.0974801	0.0854023	0.0673144	0.0487787
1.25	0.0284323	0.0478652	0.0746529	0.10356	0.122459	0.120145	0.0981212	0.068849	0.0433447
1.5	0.0208289	0.0406681	0.0733287	0.114951	0.145656	0.141749	0.106625	0.0657595	0.0357468
1.75	0.0145382	0.0328233	0.0686086	0.12271	0.168188	0.162146	0.111178	0.0597583	0.0280119
2	0.00982364	0.0255446	0.0618787	0.12709	0.189974	0.181225	0.112272	0.0523246	0.0211795

Note: $\Delta_{ij} = P_j(\theta_i) - P_j'(\theta_i')$.

A3 – Values of Δ_{ij} for $\alpha' = \alpha + .5$ and $\beta' = \beta + .4$

Discrimination Parameter α_j	Location Difference $\theta_i - \beta_j$								
	-2	-1.5	-1	-.50	0	.5	1	1.5	2
.25	0.209838	0.180626	0.13881	0.0858298	0.0254779	-0.0366199	-0.0943831	-0.14289	-0.179407
.50	0.16937	0.166617	0.146421	0.106455	0.050329	-0.011777	-0.0670872	-0.106319	-0.126847
.75	0.125209	0.14325	0.146024	0.1238	0.0749314	0.0126114	-0.0415341	-0.0732959	-0.0824953
1	0.0869699	0.11656	0.139059	0.137417	0.0991681	0.0363268	-0.0188266	-0.0463176	-0.0499357
1.25	0.0579073	0.0909571	0.127523	0.147154	0.122929	0.0591712	0.000346402	-0.0260829	-0.0283419
1.5	0.0374938	0.0688042	0.113416	0.153119	0.146114	0.0809717	0.0157081	-0.0120241	-0.0150678
1.75	0.0238368	0.0508712	0.0984051	0.155619	0.168631	0.101584	0.0273281	-0.0029612	-0.00738811
2	0.0149738	0.0369898	0.0837003	0.155091	0.190402	0.120895	0.0355237	0.00240525	-0.00318304

Note: $\Delta_{ij} = P_j(\theta_i) - P_j^{'}(\theta_i^{'})$.

Cases (II), (III), (V), and (VI) for Linear Drift of Item Difficulty and Discrimination Parameters

Case (II) – Different slope for α'

For non-zero real numbers δ and ω_j,

(a) $\alpha_j' = \delta^{-1}(\omega_j \alpha_j) = (\delta^{-1}\omega_j)\alpha_j$

(b) $\text{logit}[P_j'(\theta_i')] = (\omega_j \alpha_j)(\theta_i - \beta_j) = B_j \cdot \text{logit}[P_j(\theta_i)]$

where B_j is a *multiplicative* bias coefficient and δ is the *global* transformation parameter required to link scales.

Case (III) – Non-zero intercept and different slope for α'

For non-zero real numbers δ, τ_j, and ω_j,

(a) $\alpha_j' = \delta^{-1}(\omega_j \alpha_j + \tau_j) = (\delta^{-1}\omega_j)\alpha_j + (\delta^{-1}\tau_j)$

(b) $\text{logit}[P_j'(\theta_i')] = (\omega_j \alpha_j + \tau_j)(\theta_i - \beta_j) = B_j \cdot \text{logit}[P_j(\theta_i)] + B_j^{\theta,\beta}$

where again B_j is a *multiplicative* bias coefficient, $B_j^{\theta,\beta}$ is an *additive* bias coefficient whose absolute magnitude depends on the location difference $\theta_i - \beta_j$ and δ is the *global* transformation parameter required to link scales. This case is of course a combination of the two cases above.

<u>Case (V) – Different slope for β'</u>

For non-zero real numbers ε, δ, and λ_j,

(a) $\beta_j' = \varepsilon + \delta(\lambda_j \beta_j) = \varepsilon + (\delta\lambda_j)\beta_j$

(b) $\text{logit}[P_j'(\theta_i')] = \alpha_j(\theta_i - (\lambda_j \beta_j)) = \text{logit}[P_j(\theta_i)]^{\#}$

where ε and δ are the *global* transformation parameters required to link scales. The pound sign superscript ($^{\#}$) on the right hand side indicates that this is a transformed logit that cannot be written compactly using the original logit and bias coefficients.

<u>Case (VI) – Different slope and intercept for β'</u>

For non-zero real numbers ε, δ, κ_j, and λ_j,

(a) $\beta_j' = \varepsilon + \delta(\lambda_j \beta_j + \kappa_j) = (\varepsilon + \delta\kappa_j) + (\delta\lambda_j)\beta_j$

(b) $\text{logit}[P_j'(\theta_i')] = \alpha_j(\theta_i - (\lambda_j \beta_j + \kappa_j)) = \text{logit}[P_j(\theta_i)]^{\#} + B_j^{\alpha}$

where ε and δ are the *global* transformation parameters required to link scales. The pound sign superscript ($^{\#}$) again indicates that the first part cannot be compactly written using bias coefficients and $B_j^{\alpha} = -\alpha_j \kappa_j$ is again an *additive* bias coefficient whose magnitude depends for each item on its discrimination parameter.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM035006

I. DOCUMENT IDENTIFICATION:

Title: Bias coefficient for lack of invariance in unidimensional IRT models

Author(s): André A. Rupp, Bruno D. Zumbo

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature:	Printed Name/Position/Title: André A. Rupp	
Organization/Address: University of British Columbia, 2125 Main Mall, Vancouver, BC, V6T 1Z4	Telephone: 604-609 6691	FAX: N/A
	E-Mail Address: aarupp@interchange.ubc.ca	Date: May 8, 2003

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org